TITLE

Model Adaptation System and Method for Speaker Recognition

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention generally relates to a system and method for speaker recognition. In particular, although not exclusively, the present invention relates to speaker recognition incorporating Gaussian Mixture Models to provide robust automatic speaker recognition in noisy communications environments, such as over telephony networks and for limited quantities of training data.

Discussion of the Background Art

In recent years, the interaction between computing systems and humans has been greatly enhanced by the use of speech recognition software. However, the introduction of speech based interfaces has presented the need for identifying and authenticating speakers to improve reliability and provide additional security for speech based and related applications.

Various forms of speaker recognition systems have been utilised in such areas as banking and finance, electronic signatures and forensic science. An example of one such system is that disclosed in International Patent Application WO 99/23643 by T-Netix, Inc entitled 'Model adaptation system and method for speaker verification'. The T-Netix document describes a system and method for adapting speaker verification models to achieve enhanced performance during verification and particularly, to a sub-word based speaker verification system having the capability of adapting a neural tree network (NTN), Gaussian mixture model (GMM), dynamic time warping template (DTW), or combinations of the above, without requiring additional time consuming retraining of the models.

Another example of a speaker recognition system is disclosed in US Patent No. 6,088,699 by Maes (assigned to IBM) and is entitled 'Speech recognition with attempted speaker recognition for speaker model pre-fetching or alternative speech modelling'. Maes describes a system of identifying a speaker by text-independent comparison of an input speech signal with a stored representation of speech signals

corresponding to one of a plurality of speakers. The method of speaker recognition proposed by Maes utilises Vector Quantisation (VQ) scoring.

US Patent No. 6,411,930 by Burges (assigned to Lucent Technologies Inc.) entitled 'Discriminative Gaussian mixture models for speaker verification' discloses a method of speaker recognition that utilises a Discriminative Gaussian mixture model (DGMM). A likelihood sum of the single GMM is factored into two parts, one of which depends only on the Gaussian mixture model, and the other of which is a discriminative term. The discriminative term allows for the use of a binary classifier, such as a Support Vector Machine (SVM).

Another example of speaker recognition is discussed in US Patent No. 6,539,351 by Chen et al (assigned to IBM) and entitled 'High dimensional acoustic modelling via mixtures of compound Gaussians with linear transforms'. Chen describes a method of modelling acoustic data with a combination of a mixture of compound Gaussian densities and a linear transform. All the methods disclosed for training the model combined with the linear transform utilise the Expectation Maximization (EM) method using an auxiliary function to maximise the likelihood.

The systems described above do not provide a speaker recognition algorithm which performs reliably under adverse communications conditions, such as limited enrolment speech, channel mismatch, speech degradation and additive noise, which typically occur over telephony networks.

It would be advantageous if a system and method of speaker recognition could be provided that is robust and would mitigate the effects of adverse communications conditions, such as channel mismatch, speech degradation and noise, while also enhancing speaker model estimation.

SUMMARY OF THE INVENTION

Disclosure of the Invention

In one aspect of the present invention there is provided a method of speaker modelling, said method including the steps of:

estimating a background model based on a library of acoustic data from a plurality of speakers representative of a population of interest;

training a set of Gaussian mixture models (GMMs) from constraints provided by a library of acoustic data from a plurality of speakers representative of a population of interest and the background model;

estimating a prior distribution of speaker model parameters using information from the trained set of GMMs and the background model, wherein correlation information is extracted from the trained set of GMMs;

obtaining a training sequence from at least one target speaker;

estimating a speaker model for each of the target speakers using a GMM structure based on the maximum a posteriori (MAP) criterion.

In another aspect of the present invention there is provided a system for speaker modelling, said system including:

a library of acoustic data relating to a plurality of background speakers;

a library of acoustic data relating to a plurality of reference speakers;

a database containing training sequence(s) said training sequence(s) relating to one or more target speaker(s);

a memory for storing a background model and a speaker model for said one or more target speakers; and

at least one processor coupled to said library, database and memory, wherein said at least one processor is configured to:

- estimate a background model based on a library of acoustic data from a plurality of background speakers;
- train a set of Gaussian mixture models (GMMs) from a library of acoustic data from a plurality of reference speakers and the background model;
- estimate a prior distribution of speaker model parameters using information from the trained set of GMMs and the background model, wherein correlation information is extracted from the trained set of GMMs;
- estimate a speaker model for said one or more target speaker(s), using a GMM structure based on the maximum a posteriori (MAP) criterion, wherein the MAP criterion is a function of the training sequence and the estimated prior distribution; and
- store said background model and said speaker model in said memory.

In a further aspect of the present invention there is provided a method of speaker recognition, said method including the steps of:

estimating a background model based on a library of acoustic data from a plurality of background speakers;

training a set of Gaussian mixture models (GMMs) from a library of acoustic data from a plurality of reference speakers and the background model;

estimating a prior distribution of speaker model parameters using information from the trained set of GMMs and the background model, wherein correlation information is extracted from the trained set of GMMs;

obtaining a training sequence from at least one target speaker;

estimating a speaker model for each of the target speakers using a GMM structure based on the maximum *a posteriori* (MAP) criterion, wherein the MAP criterion is a function of the training sequence and the estimated prior distribution.

obtaining a speech sample from a speaker;

evaluating a similarity measure between the speech sample and the target speaker model and between the speech sample and the background model; and

identifying whether the speaker is one of said target speakers by comparing the similarity measures between the speech sample and said target speaker model and between the speech sample and the background model.

Other normalisations at the feature, model and score levels may also be applied to the said system.

In still yet another aspect of the present invention there is provided a system for speaker modelling and verification, said system including:

a library of acoustic data relating to a plurality of background speakers;

a library of acoustic data relating to a plurality of reference speakers;

a database containing training sequences said training sequences relating to one or more target speakers;

an input for obtaining a speech sample from a speaker;

a memory for storing a background model and a speaker model for said one or more target speakers; and

at least one processor wherein said at least one processor is configured to:

- estimate a background model based on a library of acoustic data from a plurality of background speakers;

- train a set of Gaussian mixture models (GMMs) from a library of acoustic data from a plurality of reference speakers and the background model;
- estimate a prior distribution of speaker model parameters using information from the trained set of GMMs and the background model, wherein correlation information is extracted from the trained set of GMMs;
- estimate a speaker model for said one or more target speaker(s), using a GMM structure based on the maximum *a posteriori* (MAP) criterion, wherein the MAP criterion is a function of the training sequence and the estimated prior distribution; and
- store said background model and said speaker model in said memory.
- obtain a speech sample from a speaker;
- evaluate a similarity measure between the speech sample and the target speaker model and between the speech sample and the background model;
- verify if the speaker is a target speaker by comparing the similarity measures between the speech sample and the target speaker model and between the speech sample and the background model; and
- grant access to the speaker if the speaker is verified as a target speaker.

Preferably the MAP criterion is a function of the training sequence and the estimated prior distribution.

Suitably a library of correlation information is produced from the trained set of GMMs and the estimation of prior distribution of speaker model parameters is based on the library of correlation information and the background model. Most preferably, the library of correlation information includes the covariance of the mixture component means extracted from the trained set of GMM's. A prior covariance matrix of the component means may then be compiled based on this library of correlation information.

If required, an estimate of the prior covariance of the mixture component means may be determined by the use of various methods such as maximum likelihood, Bayesian inference of the correlation information using the background

model covariance statistics as prior information or reducing the off-diagonal elements.

The library of acoustic data relating to a plurality of background speakers and the library of acoustic data relating to a plurality of reference speakers may be representative of a population of interest, including but not limited to, persons of selected ages, genders and/or cultural backgrounds.

The library of acoustic data relating to a plurality of reference speakers used to train the set of GMMs is preferably independent of the library of acoustic data used to estimate the background model, i.e. no speaker should appear in both the plurality of background speakers and the plurality of reference speakers. Most desirably, a target speaker must not be a background speaker or a reference speaker.

Preferably, the evaluation of the similarity measure involves the use of the expected frame-based log-likelihood ratio.

The background model may also directly describe elements of the prior distribution. Preferably, the present invention utilises full target and background model coupling.

The estimation of the prior distribution (in the form of the speaker model component mean prior distribution) may involve a single pass approach. Alternatively, the estimation of the prior distribution may involve an iterative approach whereby the library of reference speaker models are re-trained using an estimate of the prior distribution and the prior distribution is subsequently re-estimated. This process is then repeated until a convergence criterion is met.

The speech input for both training and testing may be directly recorded or may be obtained via a communication network such as the Internet, local or wide area networks (LAN's or WAN's), GSM or CDMA cellular networks, Plain Old Telephone System (POTS), Public Switched Telephone Network (PSTN), Integrated Services Digital Network (ISDN), various voice storage media, a combination thereof or other appropriate source.

The speaker verification and identification may further include post-processing techniques such as feature warping, feature mean and variance normalisation, relative spectral techniques (RASTA), modulation spectrum processing and Cepstral Mean Subtraction or a combination thereof to mitigate speech channel effects.

BRIEF DETAILS OF THE DRAWINGS

In order that this invention may be more readily understood and put into practical effect, reference will now be made to the accompanying drawings, which illustrate preferred embodiments of the invention, and wherein:

FIG. 1 is a schematic block diagram illustrating the background model estimation process;

FIG. 2 is a schematic block diagram illustrating the process of obtaining a component mean covariance matrix in accordance with one embodiment of the invention;

FIG. 3 is a schematic block diagram illustrating speaker model estimation for a given target speaker in accordance with one embodiment of the invention;

FIG. 4 is a schematic block diagram illustrating speaker verification in accordance with one embodiment of the present invention;

FIG. 5 is a plot of Detection Error Trade off (DET) curves according to one embodiment of the present invention; and

FIG. 6 is a plot of the Equal Error Rates (EER) according to one embodiment of the present invention.

DESCRIPTION OF EMBODIMENTS OF THE INVENTION

In one embodiment of the invention there is provided a method of speaker modelling whereby prior speaker information is incorporated into the modelling process. This is achieved through utilising the Maximum A Posteriori (MAP) algorithm and extending it to contain prior Gaussian component correlation information.

This type of modelling provides the ability to model mixture component correlations by observing the parameter variations between a selection of speaker models. In the prior art previous speaker recognition modelling work assumed that the adaptation of the mixture component means were independent of other mixture components.

With reference to FIG. 1, there is illustrated the first stage in the modelling process of one embodiment of the present invention. Estimating a background model 10 for speaker recognition may be performed in accordance with various methods, which are well known in the art. In the present case, the Expectation Maximisation (EM) algorithm is used to produce the background model. Pooled acoustic reference

data 11 relating to a specific demographic of speakers (population of interest) from a given total population is trained via the EM algorithm 12 to produce a background model 13 which is a general representation of the speech characteristics of the population of interest and is typically a large order Gaussian Mixture Model (GMM).

FIG. 2 depicts the second stage of the modelling process utilised by an embodiment of the present invention. The background model 13 is adapted utilising information from a plurality of reference speakers 21 in accordance with the Maximum A Posteriori (MAP) criterion 22. The reference speaker information within this stage of the process is composed of data samples, which represent the population of interest. However, the this reference speaker information differs from the pooled acoustic reference data 11 used to obtain the background model in that it relates to a second group of speakers from the same demographic (i.e. no sample overlap). This preserves the statistical independency of the modelling process.

Utilizing MAP estimation the reference speaker data and prior information obtainable from the background model parameters are combined to produce a library of adapted speaker models, namely Gaussian Mixture Models 23.

Using the Bayesian Inference approach, the model parameter set λ for a single model is optimized according to MAP estimation criterion given a speech utterance $\mathbf{X}$. The MAP optimization problem may be represented as follows.

$$\lambda_{MAP} = \arg\max_{\lambda} p(\mathbf{X}|\lambda)p(\lambda) \qquad \textbf{(Eq. 1)}$$

One approach is to have $p(\mathbf{X}|\lambda)$ described by a mixture of Gaussian component densities, while $p(\lambda)$ is established as the joint likelihood of w_i, μ_i and Σ_i being the weights, means and diagonal covariances of the Gaussian components respectively. The fundamental assumption specified by the prior information, without consideration of the mixture component weight effects, is that all mixture components are independent. Thus $p(\lambda)$ could be represented as the product of the joint GMM weight likelihood with the product of the individual component mean and covariance pair likelihoods as given by equation (2).

$$p(\lambda) = g(w_1, w_2, \ldots, w_N)\prod_{i=1}^{N} g(\mu_i, \Sigma_i|\Theta_i) \qquad \textbf{(Eq. 2)}$$

Here, let $g(w_1, w_2, \ldots, w_N)$ be represented as a Dirichlet distribution and $g(\mu_i, \Sigma_i | \Theta_i)$ be a Normal-Wishart density. The Dirichlet density is the conjugate prior density for the parameters of a multinomial density and the Normal-Wishart density is the prior for the parameters of the normal density.

This form of joint likelihood calculation assumes that the probability density function of the component weights is independent of the mixture component means and covariances. In addition, the joint distribution of the mean and covariance elements is independent of all other mean and covariance parameters from other Gaussians in the mixture.

Thus, the MAP solution is solved by maximizing the following auxiliary function defined by equation (3).

$$\psi(\lambda, \hat{\lambda}) \propto p(\lambda) \prod_{i=1}^{N} w_i^{c_i} \left|\Sigma_i^{-1}\right|^{\frac{c_i}{2}} \exp\left\{-\frac{c_i}{2}(\mu_i - \bar{x}_i)' \Sigma_i^{-1}(\mu_i - \bar{x}_i) - \frac{1}{2} tr(S_i \Sigma_i^{-1})\right\} \quad \textbf{(Eq. 3)}$$

where

$$c_{it} = \Pr(i | x_t, \hat{\lambda}) = \frac{\hat{w}_i \, g(x_t | \hat{\mu}_i, \hat{\Sigma}_i)}{\sum_{j=1}^{N} \hat{w}_j g(x_t | \hat{\mu}_j, \hat{\Sigma}_j)}$$

$$c_i = \sum_{t=1}^{T} c_{it}$$

$$\bar{x}_i = \sum_{t=1}^{T} \frac{c_{it} x_t}{c_i}$$

$$S_i = \sum_{t=1}^{T} c_{it} (x_t - \bar{x}_i)(x_t - \bar{x}_i)'$$

This is achieved by using the Expectation-Maximization procedure to maximize this function. Under the assumption that only the mixture component means will be adapted, the resulting EM algorithm auxiliary function is presented in equation (4)

$$\psi(\lambda, \hat{\lambda}) \propto g(\lambda) \prod_{t=1}^{N} \exp\left\{-\frac{c_i}{2}(\mu_i - \bar{x}_i)' r_i (\mu_i - \bar{x}_i)\right\} \quad \textbf{(Eq. 4)}$$

Here λ and $\hat{\lambda}$ are the new and old model estimates as a function of the mixture component means. The variable c_i is the accumulated probability count ($c_i = \sum_{t=1}^{T} c_{it}$) with $c_{it} = \frac{w_i g(x_t | \hat{\mu}_i, \Sigma_i)}{\sum_{j=1}^{N} w_j g(x_t | \hat{\mu}_j, \Sigma_j)}$) for mixture component i and $\mathbf{r}_i$ is the diagonal precision matrix for each Gaussian component i ($\mathbf{r}_i = \Sigma_i^{-1}$). The vectors μ_i and $\hat{\mu}_i$ are the ith new and old adapted Gaussian means respectively, and $\bar{x}_t = \sum_{t=1}^{T} c_{it} x_t / c_i$.

For the purposes of the present invention it is assumed that the distribution of the joint mixture component means is governed by a high dimensionality Gaussian density function. In order to represent this density, let the joint vector of the concatenated Gaussian means be represented as follows. In some works, this is described using the $vec\{\cdot\}$ operator.

$$M = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix} \qquad \text{(Eq. 5)}$$

Let the concatenated vector means have a global mean given by μ_G and a precision matrix given by $\mathbf{r}_G$. Thus, for N mixture component means, with feature dimensionality D, M is a vector of length ND, while $\mathbf{r}_G$ is an ND by ND square matrix. Thus the matrix $\mathbf{r}_G^{-1}$ is comprised of N by N sets of D by D covariance blocks (with each block identified as Σ_{ij}) between the corresponding D parameters of the ith and jth mixture component mean vectors. Given these conditions, the distribution of the concatenated means may be given in full composite form such that $g(\lambda)$ is proportional to the following.

$$g(\lambda) \propto \exp\left\{-\frac{1}{2}\left(\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_N\end{bmatrix}-\begin{bmatrix}\mu_{G1}\\ \mu_{G2}\\ \vdots\\ \mu_{GN}\end{bmatrix}\right)'\begin{bmatrix}\Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1N}\\ \Sigma_{21} & \Sigma_{22} & & \Sigma_{2N}\\ \vdots & & \ddots & \vdots\\ \Sigma_{N1} & \Sigma_{N2} & \cdots & \Sigma_{NN}\end{bmatrix}^{-1}\left(\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_N\end{bmatrix}-\begin{bmatrix}\mu_{G1}\\ \mu_{G2}\\ \vdots\\ \mu_{GN}\end{bmatrix}\right)\right\} \quad \text{(Eq. 6)}$$

Equation (6) may be given in the following symbolic compressed form

$$g(\lambda) \propto \exp\left\{-\frac{1}{2}(M-\mu_G)'\mathbf{r}_G(M-\mu_G)\right\} \quad \text{(Eq. 7)}$$

In addition, the remainder of auxiliary equation (4) must be represented in a similar matrix and vector form. The result is present in equation (8).

$$\prod_{k=1}^{N}\exp\left\{-\frac{c_i}{2}(\mu_i-\bar{x}_i)'\mathbf{r}_i(\mu_i-\bar{x}_i)\right\}=\exp\left\{-\frac{1}{2}(M-\bar{x})'\mathbf{Cr}(M-\bar{x})\right\} \quad \text{(Eq. 8)}$$

Where $\mathbf{r}=\begin{pmatrix}\Sigma_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \Sigma_2 & \mathbf{0} & \vdots\\ \vdots & \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{0} & \Sigma_N\end{pmatrix}^{-1}$

and $\mathbf{C}=\begin{pmatrix}\mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \vdots\\ \vdots & \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_N\end{pmatrix}$ with $\mathbf{C}_i=\begin{pmatrix}c_i & 0 & \cdots & 0\\ 0 & c_i & 0 & \vdots\\ \vdots & 0 & \ddots & 0\\ 0 & \cdots & 0 & c_i\end{pmatrix}=c_i\mathbf{I}$

The matrix $\mathbf{C}$ is a strictly diagonal matrix of dimension ND by ND. This matrix is comprised of diagonal block matrices $\mathbf{C}_1,\mathbf{C}_2,\ldots,\mathbf{C}_N$. Each matrix $\mathbf{C}_i$ is a D dimensional identity matrix scaled by the mixture component accumulated probability count c_i that was defined earlier.

Given this information, the equation for maximizing the likelihood can be determined. The equation in this form can be optimized (to the degree of finding a local maxima) by use of the Expectation-Maximization algorithm. This gives the following auxiliary function representation shown in equation (9).

$$\psi(\lambda,\hat{\lambda}) \propto \exp\left\{-\frac{1}{2}(M-\mu_G)' \mathbf{r}_G (M-\mu_G)\right\} \times \exp\left\{-\frac{1}{2}(M-\bar{x})' \mathbf{Cr}(M-\bar{x})\right\} \quad \textbf{(Eq. 9)}$$

Expressing this in natural logarithmic from results in equation (10).

$$\ln\psi(\lambda,\hat{\lambda}) = -\frac{1}{2}(M-\mu_G)' \mathbf{r}_G (M-\mu_G) - \frac{1}{2}(M-\bar{x})' \mathbf{Cr}(M-\bar{x}) + constant \quad \textbf{(Eq. 10)}$$

Taking the partial derivates with respect to each element of M gives

$$\frac{\partial \ln\psi(\lambda,\hat{\lambda})}{\partial M} = -2(\mathbf{Cr}+\mathbf{r}_G)M + 2(\mathbf{Cr}\bar{x} + \mathbf{r}_G\mu_G) \quad \textbf{(Eq. 11)}$$

In determining the partial derivatives, the following equalities prove useful. Here m is an arbitrary variable vector and T is a symmetric matrix (i.e. $T = T'$).

$$\frac{\partial m'\mathbf{T}}{\partial m} = \mathbf{T} \qquad \frac{\partial \mathbf{T}m}{\partial m} = \mathbf{T}' \qquad \frac{\partial m'\mathbf{T}m}{\partial m} = 2\mathbf{T}m$$

In order to locate the stationary points of the auxiliary function as expressed in equation (11), the derivative is set to zero, i.e. $\frac{\partial \ln\psi(\lambda,\hat{\lambda})}{\partial M} = 0$. This reduces the equation to the form represented in equation (12).

$$(\mathbf{Cr}+\mathbf{r}_G)M = \mathbf{Cr}\bar{x} + \mathbf{r}_G\mu_G \quad \textbf{(Eq. 12)}$$

Solving for M yields the MAP solution

$$M = (\mathbf{Cr}+\mathbf{r}_G)^{-1}(\mathbf{Cr}\bar{x} + \mathbf{r}_G\mu_G) \quad \textbf{(Eq. 13)}$$

This is reducible into the form of a weighted contribution of prior and new information.

$$M = \mathbf{a}_M\bar{x} + (\mathbf{I}-\mathbf{a}_M)\mu_G \quad \textbf{(Eq. 14)}$$

where $\mathbf{a}_M = (\mathbf{Cr}+\mathbf{r}_G)^{-1}\mathbf{Cr}$

$$(\mathbf{I}-\mathbf{a}_M) = (\mathbf{Cr}+\mathbf{r}_G)^{-1}\mathbf{r}_G$$

Now given that the global mean μ_G is set to the concatenated background model means, the factor $\mathbf{a}_M$ contains information relating to the proportion of new to old information contained in the background model that is to be included in the adaptation process.

Now that the adaptation equation is capable of handling the prior correlation information within the MAP adaptation framework one method for determining the global correlation components is the Maximum Likelihood criterion. The Maximum Likelihood criterion estimates the covariance matrix through the parameter analysis of a library of Out-Of-Set (OOS) speaker models. If the correlation components describe the interaction between the mixture mean components appropriately, the adaptation process can be controlled to produce an optimal result. The difficulty with the data based approach is the accurate estimation of the unique parameters in the ND by ND covariance matrix. For a complete description of the matrix, at least ND+1 unique samples are required to avoid a rank deficient matrix or density function singularity. This implies that at least ND+1 speaker models are required to satisfy this constraint. This requirement alone can be prohibitive in terms of computation and speech resources. For example, a 128 mode GMM with 24 dimensional features requires at least 3073 well-trained speaker models to calculate the prior information.

The Maximum Likelihood solution involves finding the covariance statistics using only the out-of-set speaker models. So, if there are s^{OOS} out-of-set models trained from a single background model with the concatenated mean vector extracted from the jth model given by, μ_j^{OOS} the covariance matrix estimate, Σ_G^{ML} , is simply calculated with equation (15). If the estimate for the mean μ_G^{ML} is known, then equation (16) need not be used. Such an example is where the background component means are substituted for μ_G^{ML} .

$$\Sigma_G^{ML} = \frac{1}{s^{oss}-1}\sum_{j=1}^{s^{OOS}}\left(\mu_j^{OOS}-\mu_G^{ML}\right)\left(\mu_j^{OOS}-\mu_G^{ML}\right)' \qquad \textbf{(Eq. 15)}$$

$$\text{with} \quad \mu_G^{ML} = \frac{1}{s^{OOS}}\sum_{j=1}^{s^{OOS}}\mu_j^{OOS} \qquad \textbf{(Eq. 16)}$$

Unfortunately, if there are insufficient models to represent the covariance matrix, the matrix becomes rank deficient and no inverse can be determined. This difficulty of a rank-deficient covariance matrix is shared with subspace adaptation approaches such as "eigenvoice" analysis that are applied in both speech and speaker recognition. This difficulty may be resolved through a number of methods described below, that are also applicable to eigenvoice analysis.

One method involves Principal Component Analysis (PCA). This approach involves decomposing the matrix representation into its principal components. Once the principal components have been extracted, they may be used in conjunction with (empirical, data-derived or other) diagonal covariance information for adaptation. Restricting adaptation solely to this lower dimensional principal component subspace likewise restricts the capability for adapting model parameters outside the subspace. This causes performance degradation for larger quantities of adaptation data, which may be alleviated by using a combined approach. Ideally, a technique that can exploit some of the significant principal components of variation information with other adaptation statistics may operate robustly for both short and lengthy training utterances. In this manner, the principal components may restrict the adaptation to a subspace for small quantities of speech and will converge to the maximum likelihood solution for larger recordings.

Another solution for avoiding the generation of a singular covariance matrix, but not necessarily limited to this, is to reduce the magnitude of the non-diagonal covariance components. This approach allows the inverse of the matrix to be determined. It also permits the covariance matrix to allow adaptation of the target model parameters outside the adaptation subspace defined by the OOS speaker variations. The covariance estimation, given that the global mean is known, is performed using equation (17). Here $diag\{\cdot\}$ represents the diagonal covariance matrix and ξ_d is generally a small number near zero but between zero and one.

$$\Sigma_G = \xi_d diag\left\{\Sigma_G^{ML}\right\} + (1-\xi_d)\Sigma_G^{ML} \qquad \textbf{(Eq. 17)}$$

Another possible method for determining the global correlation components is Bayesian adaptation of the covariance and (if required) the mean estimates by combining the old estimates from the background model with new information from a

library of reference speaker models. The reference speaker data library is comprised of s^{OOS} out-of-set speaker models represented by the set of concatenated mean vectors, $\{\mu_j^{OOS}\}$. In addition, the old mean and covariance statistics are given by μ_G^{old} and Σ_G^{old} respectively.

$$\Sigma_G^{adapt} = \xi E\left\{\mu_j^{OOS}\mu_j^{OOS'}\right\} + (1-\xi)\left(\Sigma_G^{old} + \mu_G^{old}\mu_G^{old'}\right) - \mu_G^{adapt}\mu_G^{adapt'} \qquad \text{(Eq. 18)}$$

$$\mu_G^{adapt} = \xi\mu_G^{ML} + (1-\xi)\mu_G^{old} \qquad \text{(Eq. 19)}$$

$$\text{with} \quad E\left\{\mu_j^{OOS}\mu_j^{OOS'}\right\} = \frac{1}{s^{OOS}}\sum_{j=1}^{s^{OOS}}\mu_j^{OOS}\mu_j^{OOS'} \qquad \text{(Eq. 20)}$$

$$\xi = \frac{s^{OOS}}{s^{OOS} + s^{old}} \qquad \text{(Eq. 21)}$$

If the global mean vector estimate is known then $\mu_G^{adpt} = \mu_G^{old} = \mu_G^{ML}$. One estimate may be to set these parameters to the background model mean vector μ_G^{BM}. In the instance that the mean of the Gaussian distribution is known, and only the covariance information is adapted, the adapted covariance becomes equation (22).

$$\Sigma_G^{adapt} = \xi\Sigma_G^{ML} + (1-\xi)(\tau\mathbf{r})^{-1} \qquad \text{(Eq. 22)}$$

The prior estimate of the global covariance, according to standard adaptation techniques, is given by $(\tau\mathbf{r})^{-1}$ while the new information is supplied by the covariance statistics determined from the collection of OOS speaker models. The hyperparameter τ is the relevance factor for the standard adaptation technique and the matrix $\mathbf{r}$ is the diagonal concatenation of the Gaussian mixture component precision matrices. The variable ξ is a tuning factor that represents how important the sufficient statistics, which are derived from the ML trained OOS models, are relative to the UBM based diagonal covariance information. Now, if the OOS model derived covariance information is unreliable, ξ should reduce to 0. In this case, the adaptation equation then resolves into the basic coupled mixture component mean adaptation system i.e. $M = (\mathbf{Cr} + \mathbf{r}_G)^{-1}(\mathbf{Cr}\bar{x} + \mathbf{r}_G\mu_G)$ becomes $M = (\mathbf{Cr} + \tau\mathbf{I})^{-1}(\mathbf{C}\bar{x} + \tau\mu_G)$.

However, as the value of ξ increases, the emphasis on using covariance information derived from the multiple OOS speaker models is increased. The strength of MAP estimation of the covariance statistic is that the adapted covariance matrix will not be rank deficient provided the old covariance information is of full rank and ξ is less than 1.

Thus in accordance with the EM algorithm with the MAP criterion the reference speaker data $\mathbf{X}^{OOS}$ 21 is utilised to adapt the background model for each speaker contained in the reference speaker data library to form a set of adapted speaker models in the form of GMM's 23.

The covariance statistics of the component means are then extracted from this adapted library of models 24 using standard techniques, see equation 15. The resultant of this extraction is the formation of a component mean covariance (CMC) matrix 25. The CMC matrix may then be used in conjunction with the background model 13 to estimate the prior distribution for controlling the target speaker adaptation process.

With reference to FIG. 3, there is illustrated the third stage of the modelling process utilised by the present invention. The background model 13 and the CMC matrix 25 are combined to estimate the prior distribution 31 for the set of component means.

Alternatively, the CMC matrix may be used in further iterations of reference speaker model training, in this instance the CMC data is fed back to re-train the reference speaker data with the background model, and then re-estimating the CMC matrix. This joint optimization process allows for variations of the mixture components to not only become dependent on previous iterations but also on other components further refining the MAP estimates. Several criteria may be used for this joint optimization of the reference models with the prior statistics, such as the maximum joint *a posteriori* probability over all reference speaker training data, eg.

$$\Sigma_G^{MJAP} = \arg\max_{\Sigma_G} \sum_i \log \max_{\lambda_i} p(\mathbf{X}_i|\lambda_i) p(\lambda_i|\Sigma_G) \qquad \textbf{(Eq. 23)}$$

A training sequence is acquired for a given target speaker either directly or from a network 32. For normal training of speaker recognition models at least 1 to 2 minutes of training speech is required. This training sequence and the prior

distribution estimate 31 are then utilised in conjunction with the MAP criterion as derived in the above discussion to estimate a speaker model for a given target speaker 34.

The target speaker model produced in this instance incorporates model correlations into the prior speaker information. This enables the present invention to handle applications where the length of the training speech is limited.

FIG. 4 illustrates one possible application of the present invention namely that of speaker verification 40. A speech sample 41 is obtained either directly or from a network. The sample is compared against the target model 43 and the background model 42 to produce similarity measures for the sample against the target and background models. The similarity measure is preferably calculated using the expected log likelihood. When comparing the likelihood between classes the likelihood ratio may be treated as independent of the prior target and impostor class probabilities $P(\lambda_{tar})$ and $P(\lambda_{non})$. The LR statistic is expressed as:

$$LR(x_t) = \frac{p(x_t|\lambda_{tar})}{p(x_t|\lambda_{non})} \qquad \textbf{(Eq. 24)}$$

For ease of mathematically manipulating the solution the logarithm is taken, resulting in the *Log Likelihood Ratio* (LLR) which is given as:

$$LLR(x_t) = \log p(x_t|\lambda_{tar}) - \log p(x_t|\lambda_{non}) \qquad \textbf{(Eq. 25)}$$

If the likelihoods are in fact probability densities, the likelihood ratio of a single observation, may be used to determine the target speaker probability given that the sample was taken from either the target or non-target speaker distributions.

$$P(\lambda_{tar}|x_t) = \frac{LR(x_t)P(\lambda_{tar})}{LR(x_t)P(\lambda_{tar}) + P(\lambda_{non})} \qquad \textbf{(Eq. 26)}$$

Given T observations, assumed independent and identically distributed, $\mathbf{X} = (x_1, x_2, \ldots, x_T)$, the ratio of the joint likelihoods in log form is given.

$$LLR(\mathbf{X}) = \sum_{t=1}^{T} \log p(x_t|\lambda_{tar}) - \log p(x_t|\lambda_{non}) \qquad \text{(Eq. 27)}$$

In practical applications, this estimate for a target speaker model figure of merit is not a robust measure, since the observations are not independent or identically distributed and also that there is a dependence between the background model and the coupled target models. A more robust measure for speaker verification is the expected log-likelihood ratio measure given by equation 28. This measure is typically used in forensic casework applications and is typically compensated for environmental effects through score normalisation.

$$E[LLR(x_t)] = E[\log p(x_t|\lambda_{tar}) - \log p(x_t|\lambda_{non})] \qquad \text{(Eq. 28)}$$

$$= \frac{1}{T}\sum_{t=1}^{T}(\log p(x_t|\lambda_{tar}) - \log p(x_t|\lambda_{non})) \qquad \text{(Eq. 29)}$$

A similarity measure is then calculated in the above manner for the acquired speech sample 41 compared with the background model 42 and for the acquired speech sample compared with the speaker model of the target person 43. These measures are then compared 44 in order to determine if the speech sample is from the target person 45.

To demonstrate the effect of including correlation information, the present invention will be discussed with reference to FIG. 5 which represents the speaker detection performance of one embodiment of the present invention.

In this instance, a fully coupled target and background model structure was adapted using the above-described approach. Here, model coupling refers to the target model parameters being derived ~~from~~ a function of the training speech and the background model parameters. In the limit sense when there is no training speech the target speaker model is represented as the background model. The embodied system also utilised a feature warping parameterization algorithm and performed scoring of a test segment via the expected log-likelihood ratio test of the adapted target model versus the background model.

The system evaluation was based on the NIST 2000 and 1999 Speaker Recognition Databases. Both databases provide approximately 2 minutes of speech

for the modelling of each speaker. The NIST 2000 database represented a demographic of 416 male speakers recorded using electret handsets. The information of the 2000 database was used to determine the correlation statistics. While the first 5 and 20 seconds of speech per speaker in the 1999 database was used as the training samples.

Detection Error Trade-off (DET) curves for the system are shown in FIG. 5, the system curves are based on 20 second lengths of speech for a set of male speakers processed according to the extended MAP estimation condition, and whereby the number of out-of-set (OOS) speakers was increased for each estimation of the covariance matrix statistics. The selection of OOS speakers involved using 20, 50, 100, 200 and 400 speakers. The result for the baseline background model is also identified in the plot. Because the number of OOS speakers is less than the number of rows or columns in the matrix, the matrix is singular. To avoid this problem, the non-diagonal components of the covariance matrix are deemphasized by 0.1%. It is clear from FIG. 5 that utilising the correlation information in the modelling process yields a continued increase in performance for an increasing number of OOS speakers used in estimation of the covariance matrix. It is important to note that the number of speakers is significantly below the minimum of 3073 speakers required for a non-singular matrix estimate without the need of deemphasizing the non-diagonal covariance components. Ideally, the evaluation requires the number of OOS speakers to be an order of magnitude more. However, the improvement in performance by using the correlation information in the modelling process is apparent from FIG. 5.

FIG. 6 illustrates a plot of equal error rate performances for the 20-second training utterances and for 5-second utterances for the system of FIG. 5. For 5 seconds of training speech, using the correlation information, the EER is reduced from 28.8% for 20 speakers to 20.4% for 400 speakers. Correspondingly, the 20 second results indicated an improving performance trend of 24.3% EER for 20 speakers down to 16.6% EER for 400 speakers. In both instances the background model based system performance exceeded that of the best covariance approximation system giving a 14.8% EER. However it is to be noted that background model based system error rates would be outperformed by the covariance prior estimate system if more OOS speakers were available as the

background model baseline covariance matrix is far from becoming an accurate estimate of the true covariances.

It is to be understood that the above embodiments have been provided only by way of exemplification of this invention, and that further modifications and improvements thereto, as would be apparent to persons skilled in the relevant art, are deemed to fall within the broad scope and ambit of the present invention defined in the following claims.